Published online 21 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D549–D553***
doi:10.1093/nar/gkr1049

ProtChemSI: a network of protein–chemical structural interactions

Olga V. Kalinina[1,2], Oliver Wichmann[1], Gordana Apic[1,3] and Robert B. Russell[1,*]

[1]Cell Networks, BioQuant, University of Heidelberg, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany, [2]Institute for Information Transmission Problems, RAS, Bolshoi Karenty pereulok 19, Moscow, 127994, Russia and [3]Cambridge Cell Networks Ltd, St John's Innovation Centre, Cowley Road, Cambridge CB4 OWS, UK

Received August 15, 2011; Revised October 4, 2011; Accepted October 25, 2011

ABSTRACT

Progress in structure determination methods means that the set of experimentally determined 3D structures of proteins in complex with small molecules is growing exponentially. ProtChemSI exploits and extends this useful set of structures by both collecting and annotating the existing data as well as providing models of potential complexes inferred by protein or chemical structure similarity. The database currently includes 7704 proteins from 1803 organisms, 11 324 chemical compounds and 202 289 complexes including 178 974 predicted. It is publicly available at http://pcidb.russelllab.org.

INTRODUCTION

Protein–chemical interactions are most often not considered in the context of three-dimensional (3D) structures. Most databases, such as DrugBank (1) or STITCH (2) will refer to 3D structures but do not exploit them beyond reporting that a structure for a drug–protein interaction is known. Other databases, such as Binding MOAD (3), PDBbind (4) and BindingDB (5), focus on collecting protein–ligand complexes, but report only those that are experimentally resolved. However, the current network of protein–chemical interactions derived from 3D structures is a rich source of information and provides many possibilities to suggest new protein–chemical interactions.

Recently, we published a method to predict novel protein–chemical interactions using superimposition of known 3D structures (6). The underlying principle is that if two proteins share a common ligand, and the first protein is known to bind a second ligand, the 3D structures of protein–ligand complexes can be superimposed to build a model that can be used to evaluate a complex of the second protein with that second ligand (Figure 1, lower). Here we present ProtChemSI, a database providing these computed complexes. The database also contains known structures of protein–chemical complexes, and several other predicted complexes. Specifically, we also construct models for all interactions with molecules similar to known interaction partners of a protein or a chemical of interest (Figure 1, explained in detail below), and provide a method to traverse the network of interactions to identify possibilities for building a structural model of any protein chemical pair of interest (Figure 2).

Being primarily based on structural interactions, ProtChemSI has little overlap with other databases for protein–chemical interactions, such as DrugBank (1), STITCH (2) and ChEMBL (7) (Table 1). Theoretically, protein–chemical interactions viewed as a network provide a possibility to construct a model of a complex of any given protein and chemical, superimposing molecules along the path that connects them. ProtChemSI implements a routine to construct and evaluate these models on user demand, so the total number of theoretically possible models in ProtChemSI is very large and impossible to quantify. However, including first-order models (i.e. where we consider interactions no more than two-steps away in the network), we have a total of 23 315 known complexes, and predictions, where 65 502 are modeled by obvious homology, 18 917 are modeled by obvious chemical similarity and 94 555 are modeled by superimpositions as detailed in our original study (6).

FUNCTIONS OF THE DATABASE

ProtChemSI is intended for those interested in structural details of interactions between proteins and small molecules. It provides details at two levels of certainty: first, it lists all experimentally resolved 3D structures involving the query protein or chemical; second, it constructs a number of models as detailed below.

The workflow of the model construction is schematically represented in Figure 1. For a query protein, models of the following complexes are constructed: (i) with

*To whom correspondence should be addressed. Tel: +49 6221 54 51 362; Fax: +49 6221 54 51 486; Email: robert.russell@bioquant.uni-heidelberg.de
Present address:
Olga V. Kalinina, Max-Planck-Institut für Informatik, Campus E1 4, 66123 Saarbrücken, Germany.

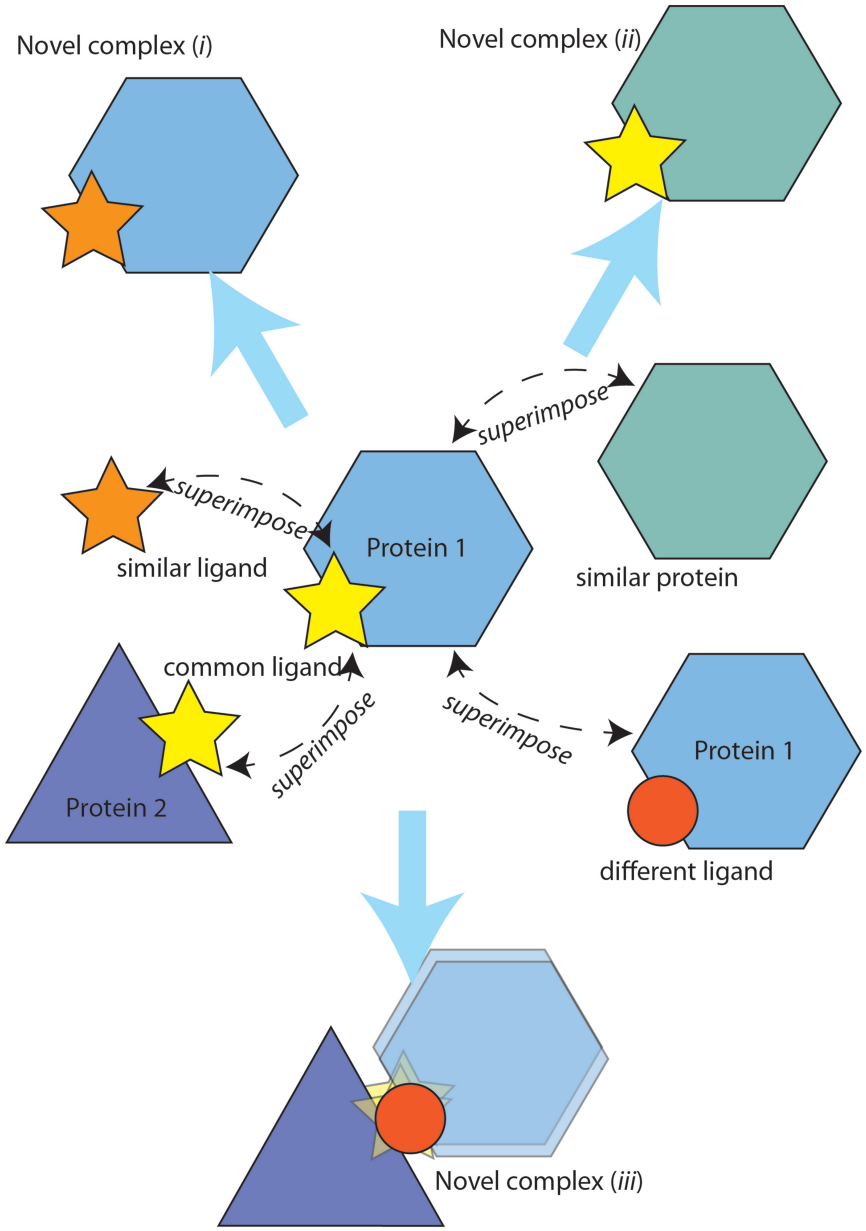

Figure 1. Schematic representation of the construction of the models of protein–ligand complexes. The models are constructed using superimposition of similar chemicals (top-left corner), superimposition of similar protein (top-right corner) or double superimposition of identical proteins and chemicals in the case when two proteins share a common ligand and one of them has another ligand (bottom).

chemicals that are similar to experimentally observed binding ligands (Figure 1, upper left); (ii) with chemicals that bind to proteins similar to the query (Figure 1, upper right); (iii) with chemicals that bind to another protein that shares at least one binding ligand with the query (Figure 1, bottom). The models for chemical queries are constructed analogously. All the models are scored according to statistics described in (6). Briefly, a number of physical and chemical parameters of the complex are combined to give a score from 0 to 7, with 7 being the best possible model. Models with a $P \leq 0.05$ (which corresponds to a score of 5.6), i.e. having a $\leq$5% chance of being randomly generated, are highlighted for the user.

Another application within ProtChemSI is a shortest-path and superimposition functionality. When given a protein and a chemical, this feature attempts to construct a complex of them by first finding the shortest path between the molecules through the network of experimentally determined structures. It then sequentially superimposes them using the similar components of two adjacent complexes as templates, similar to Figure 1, to obtain a complex of the two components of interest. Then it evaluates them and reports as described above.

An example of this shortest path reconstruction is presented in Figure 2. We can construct a complex of the well-documented (1) interaction between the human

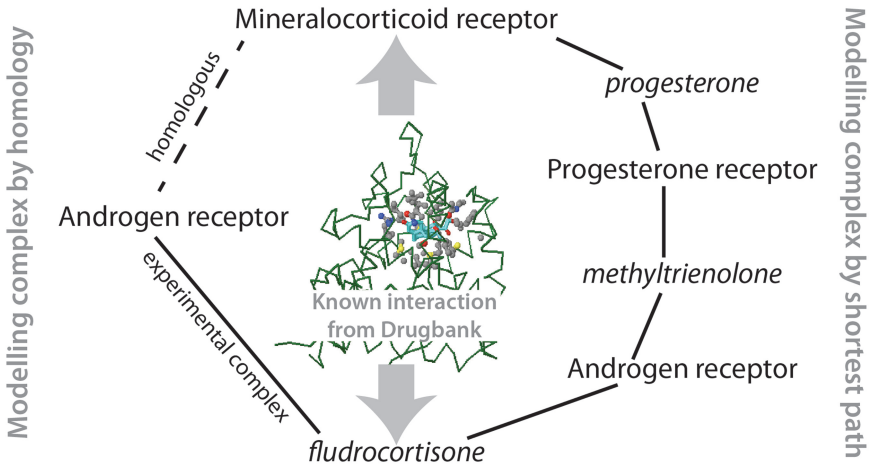

Figure 2. Example of reconstruction of a known interaction. Mineralocorticoid receptor is documented as a target for fludrocortisone in the DrugBank (1). ProtChemSI reconstruct the complex corresponding to this interaction either using superimposition of similar proteins (left part) of a chain of superimpositions of protein–ligand complexes that constitute the shortest path in the network between Mineralocorticoid receptor and fludrocortisone (right part). The main chain of the protein is shown in green, the ligand is shown in sticks mode. The contacting atoms of the protein are shown in ball-and-stick mode and colored by atom type with gray carbons.

Table 1. Overlap with other sources for protein-chemical interactions

	Number of proteins	Number of chemicals	Number of protein–chemical direct complexes/ interactions	Number of models
DrugBank	2311	1944	2770	1010
STITCH	3091	2661	1861	957
ChEMBL	1006	2083	1324	3038

Mineralocorticoid receptor and fludrocortisone in two ways. First, we get a complex by superimposing the 3D structure of the Mineralocorticoid receptor onto the structure of the homologous human Androgen receptor complexed with fludrocortisone, with a highly significant score of 6.43. The second route is to do a chain of superimpositions of the complexes of Mineralocorticoid receptor with progesterone, progesterone with Progesterone receptor, Progesterone receptor with methyltrienolone, methyltrienolone with Androgen receptor and Androgen receptor with fludrocortisone, also giving a highly significant score of 6.42.

CONTENT OF THE DATABASE

The database of ProtChemSI rests on two main concepts: molecules and links. Molecules are proteins and chemicals [defined as anything with a corresponding entry in PubChem (8)] that appear in the Protein Data Bank (9) to be in contact in an X-ray or NMR resolved 3D structure. Links represent either complexes of a protein and a chemical or a similarity relationship between two proteins or two chemicals. Two proteins are considered similar if they are >30% identical in a BLAST (10) search with E-value < 0.01, and two chemicals, if the Tanimoto score [using PubChem fingerprints (11)] between them is ≥ 0.9.

Chemicals with less than five heavy atoms are currently ignored. We also filter out solvents, buffer components and other non-specifically binding ligands. To do this, we inspected all ligands with more than 20 linked proteins manually and filtered those belonging to one of these categories. Chlorophyll, heme and other porphyrins were also removed, as they represent a very specific case of protein–ligand interaction that tend to obscure more specific interactions made by smaller ligands.

For each pair of instances of the same molecule, or similar molecules, a transformation matrix bringing them into the same frame of reference is also stored. The models that are being reported are not stored explicitly, but are calculated on-demand. The recently computed models are cached and kept for 2 weeks.

The database is updated monthly. The flat-files of the current release can be downloaded from the website.

WEB INTERFACE

There are two main routes to interact with ProtChemSI: browsing the database, or searching by protein name, UniProt ID, protein sequence, ligand name, PubChem ID or SMILES string. For any given protein or chemical, a number of features are shown: (i) known binding chemicals/proteins from the experimentally resolved 3D structures; (ii) possible interacting chemicals/proteins that are similar to those bound by the query and form plausible models based on the statistics described above; (iii) possible interacting chemicals/proteins that bind molecules similar to the query and form plausible models; and (iv) possible interacting chemicals/proteins from all one-step models as represented in Figure 1. It is possible to select a second molecule in the network and attempt to construct a model of the complex of the two, using the chain of superimpositions, as described above.

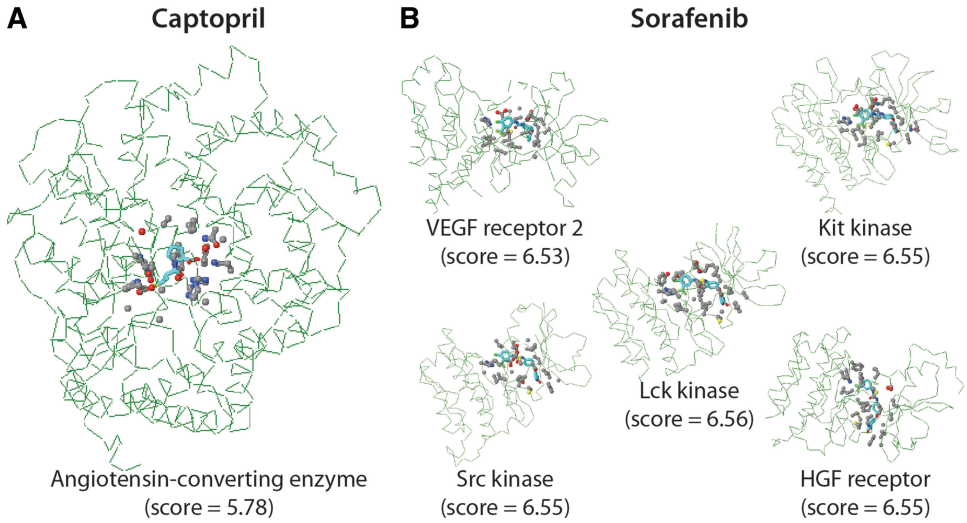

Figure 3. (**A**) Model of Captopril with its natural target, human angiotensin I-converting enzyme (model score 5.78, highly significant). (**B**) Model of Sorafenib with its known (VEGF receptor 2, Kit kinase) and potential (Lck and Src kinases, hepatocyte growth factor receptor) targets. Representation and coloring scheme same as in Figure 2.

USE CASES

The database can provide interesting suggestions of new compounds binding to proteins, particularly when several protein–complex structures have already been determined. For instance, when one searches for interactions involving human carbonic anhydrase 2 (CAH2), a protein with many known chemical inhibitors, the server first reports 133 complexes of known structure, 11 complexes predicted based on homologous proteins (of which 3 are significant), 4 predictions (none significant) based on similar chemicals and 44 predictions (2 significant) based on protein/ligand superimpositions as described in the original paper. Among these are several somewhat obvious predictions (e.g. where a close homolog is used to make the prediction), but also some intriguing suggestions such as hydroxyaminovaline, which binds to Collagenase 3 (MMP13) which in turn shares the bound chemical acetohydroxamate with CAH2. This prediction is structurally plausible, with the additional propyl group in hydroxaminovaline relative to acetohydroxamate fitting nicely into a hydrophobic pocket, however to our knowledge this compound is not currently known to affect CAH2.

The database can also provide structures for known protein–chemical interactions lacking an experimental structure. We tested this by inspecting all interactions within Drugbank (1). There are 2785 experimentally resolved 3D structures of DrugBank drug–target interactions within our database. Additionally, we construct 1100 models involving drugs and their targets. For example, for a specific inhibitor of angiotensin I-converting enzyme (ACE), Captopril, which is used to treat hypertension, only its complex with the Drosophila ortholog is resolved (PDB code 1J37). Despite low identity to the human protein (45%), we are able to reconstruct confident models with its cognate human target (Figure 3A). Another interesting example is Sorafenib, an anti-cancer drug that is reported to bind a number of protein kinases. Using resolved structures with two of them, MAPK14 (PDB codes 3GSC, 3HEG) and B-raf (PDB codes 1UWJ, 1UWH), we reconstruct models with known targets VEGF receptor 2 and Kit kinase, as well as with a number of novel potential targets, Lck and Src kinases and hepatocyte growth factor receptor (Figure 3B).

It is also possible to find new suggestions for well-studied chemicals such as metabolites. For instance, using cholesterol as a query, one obtains, among others, its model with an orphan receptor LXR-β (using similarity to ROR-γ) or with a yeast oxystrerol-binding protein KES1 (using 1-step superimposition via NPC1 and 25-hydroxycholesterol). Even some of the most well-studied drugs have some intriguing findings, for instance, Aspirin is predicted with confidence to bind to β-galactosidase from *Escherichia coli*, which is not the case for other non-steroidal anti-inflammatory agents of the same class, such as diclofenac or acetaminophen/paracetamol, a result which agrees broadly with a study of the effect of these chemicals on the human enzyme (12,13).

CONCLUSION

The database provides a view of protein–chemical interactions from a structural perspective. Although these data are currently growing exponentially, the structural angle is often missing from other resources. We provide structural models for well-known interactions and provide putative models for millions of others. The database provides an excellent complement to existing tools to study protein–chemical interactions. Due to the nature of the method behind the database, both the number and quality of models in ProtChemSI will increase as the set of experimentally determined complexes grows. We anticipate that the database will play as important a role in the study of

protein–ligand interactions as homology modeling plays in structural biology in general.

ACKNOWLEDGEMENTS

The authors are grateful to Dr Matthew J. Betts and Dr Chad A. Davis for the technical support and the critical reading of the manuscript.

FUNDING

This work was funded under the Excellence Initiative "Cell Networks" from the German Science Ministry (DFG). Funding for open access charge: DFG.

Conflict of interest statement. None declared.

REFERENCES

1. Wishart,D.S., Knox,C., Guo,A.C., Cheng,D., Shrivastava,S., Tzur,D., Gautam,B. and Hassanali,M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.
2. Kuhn,M., Szklarczyk,D., Franceschini,A., Campillos,M., von Mering,C., Juhl Jensen,L., Beyer,A. and Bork,P. (2010) STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.*, **38**, D552–D556.
3. Hu,L., Benson,M.L., Smith,R.D., Lerner,M.G. and Carlson,H.A. (2005) Binding MOAD (Mother Of All Databases). *Proteins*, **60**, 333–340.
4. Wang,R., Fang,X., Lu,Y. and Wang,S. (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.*, **47**, 2977–2980.
5. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
6. Kalinina,O.V., Wichmann,O., Apic,G. and Russell,R.B. (2011) Combinations of protein-chemical complex structures reveal new targets for established drugs. *PLoS Comp. Biol.*, **7**, e1002043.
7. de Matos,P., Alcántara,R., Dekker,A., Ennis,M., Hastings,J., Haug,K., Spiteri,I., Turner,S. and Steinbeck,C. (2010) Chemical entities of biological interest: an update. *Nucleic Acids Res.*, **38**, D249–D254.
8. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
9. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
10. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. National Center for Biotechnology Information. 'PubChem Substructure Fingerprint v1.3.' PubChem Data Specification Directory, 1 May 2009. ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt (8 November 2011, date last accessed).
12. Bode-Böger,S.M., Martens-Lobenhoffer,J., Täger,M., Schröder,H. and Scalera,F. (2005) Aspirin reduces endothelial cell senescence. *Biochem. Biophys. Res. Commun.*, **334**, 1226–1232.
13. Hu,Z., Zhang,F., Yang,Z., Zhang,J., Zhang,D., Yang,N., Zhang,Y. and Cao,K. (2008) Low-dose aspirin promotes endothelial progenitor cell migration and adhesion and prevents senescence. *Cell Biol. Int.*, **32**, 761–768.